Exercise 1: Connection between MLE and ERM

Suppose we are facing a regression task, i.e., $\mathcal{Y} = \mathbb{R}$, and the feature space is $\mathcal{X} \subseteq \mathbb{R}^p$. Let us assume that the relationship between the features and labels is specified by

$$y = m^{-1}\left(m(f_{\text{true}}(\mathbf{x})) + \epsilon\right), \tag{1}$$

where $m : \mathbb{R} \to \mathbb{R}$ is a continuous strictly monotone function with m^{-1} being its inverse function, and the errors are Gaussian, i.e. $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In particular, for the data points $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$ it holds that

$$y^{(i)} = m^{-1}\left(m(f_{\text{true}}(\mathbf{x}^{(i)})) + \epsilon^{(i)}\right), \tag{2}$$

where $\epsilon^{(1)}, \ldots, \epsilon^{(n)}$ are iid with distribution $\mathcal{N}(0, \sigma^2)$.

Disclaimer: We assume in the following that $m(y)$ and $m(f(\mathbf{x}))$ is well-defined for any $y \in \mathcal{Y}$, $f \in \mathcal{H}$ and $\mathbf{x} \in \mathcal{X}$.

(a) How can we transform the labels $y^{(1)}, \ldots, y^{(n)}$ to "new" labels $z^{(1)}, \ldots, z^{(n)}$ such that $z^{(i)} \mid \mathbf{x}$ is normally distributed? What are the parameters of this normal distribution?

(b) Assume that the hypothesis space is

$$\mathcal{H} = \{f(\cdot \mid \boldsymbol{\theta}) : \mathcal{X} \to \mathbb{R} \mid f(\cdot \mid \boldsymbol{\theta}) \text{ belongs to a certain functional family parameterized by } \boldsymbol{\theta} \in \Theta\},$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$ is a parameter vector, which is an element of a **parameter space** Θ. Based on your findings in (a), establish a relationship between minimizing the negative log-likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \ldots, (\mathbf{x}^{(n)}, z^{(n)})$ and empirical loss minimization over $\mathcal{H}$ of the generalized L2-loss function of Exercise sheet 1, i.e., $L\left(y, f(\mathbf{x})\right) = \left(m(y) - m(f(\mathbf{x}))\right)^2$.

(c) In many practical applications such as biology, medicine, physics or social sciences one often observed statistical property is that the label y given a feature $\mathbf{x}$ follows a *log-normal distribution*[1]. Note that we can obtain such a relationship by using $m(x) = \log(x)$ above. In the following we want to consider the conjecture of the Scottish physician James D. Forbes, who conjectured in the year 1857 that the relationship between the air pressure (in inches of mercury) y and the boiling point of water x (in degrees Farenheit) is given by

$$y = \theta_1 \exp(\theta_2 x + \epsilon),$$

for some specific values $\theta_1 \in \mathbb{R}_+, \theta_2 \in \mathbb{R}$ and some error term ϵ (of course, we assume that this error term is stochastic and normally distributed).

- What would be a suitable hypothesis space $\mathcal{H}$ if this conjecture holds?
- The dataset `forbes` in the R-package `MASS` contains 17 different observations of y and x at different locations in the Alps and Scotland, i.e., the data set is $(x^{(i)}, y^{(i)})_{i=1}^{17}$. Analyze whether his conjecture was reasonable by using the following code snippet:

```
#' @param X the feature input matrix X
#' @param y the outcome vector y
#' @param theta parameter vector for the model (2-dimensional)

# Load MASS and data set forbes
library(MASS)
data(forbes)
attach(forbes)
```

[1] The Wikipedia article on the log-normal distribution has quite a large part about the occurrence of the log-normal distribution.

```r
# initialize the data set
X = cbind(rep(1,17),bp)
y = pres

#' function to represent your models via the parameter vector theta = c(theta_1, theta_2)
#' @return a predicted label y_hat for x
f <- function(x, theta){

  # >>> do something <<<

  return(y_hat)

}

#' @return a vector consisting of the optimal  parameter vector
optim_coeff <- function(X,y){

  # >>> do something <<<

  return(theta)

}

  # >>>  Do something here to check Forbes' conjecture <<<
```

Hint: As a sanity check whether your function to find the optimal coefficients work, it should hold that $\hat{\theta}_1 \approx 0.3787548$ and $\hat{\theta}_2 \approx 0.02062236$.